

BAB I

PENDAHULUAN

A. Latar Belakang

SMS atau *Short Message Service* merupakan salah satu layanan teknologi yang digunakan dalam melakukan pertukaran pesan melalui perangkat mobile seperti telepon genggam. SMS adalah cara penyimpanan dan langkah maju dalam melakukan transmisi pesan ke dan dari ponsel atau telepon genggam. Pesan dari ponsel pengirim disimpan dalam pusat pesan singkat (SMS Center) yang kemudian diteruskan ke perangkat tujuan. Setiap pesan singkat tidak lebih dari 160 karakter, sementara karakter ini dapat berupa teks (alfanumerik) atau biner pesan pendek Non-Teks (Katankar and Thakare, 2010).

SMS tidak memiliki *header* pada setiap pesan yang dikirim seperti pada email. Pesan teks juga penuh dengan kata singkatan dan memiliki lebih sedikit bahasa formal dari apa yang akan diharapkan dari email. Semua faktor ini dapat mengakibatkan penurunan serius dalam kinerja banyak algoritma penyaringan email spam terkenal apabila diterapkan pada pesan teks singkat.

Sebagian besar aplikasi yang menggunakan metode Linear Vector Quantization (LVQ) dalam aplikasinya digunakan untuk menyelesaikan masalah yang berhubungan dengan klasifikasi. Salah satu fitur yang paling menakjubkan dari LVQ adalah algoritma ini mengambil sedikit vektor untuk hasil klasifikasi yang sangat baik (Patra, Dash and Tripathy, 2013).

Learning Vector Quantization menumbuhkan satu set neuron dan memilih neuron terbaik untuk setiap tugas klasifikasi dan melatih neuron – neuron untuk meningkatkan akurasi mereka. LVQ sangat cocok untuk tugas – tugas klasifikasi teks, dan telah diterapkan untuk masalah klasifikasi SPAM dengan hasil unggul baik dengan Bayesian dan berbagai bentuk lain dari jaringan syaraf tiruan (Hasan Alkahtani, Paul Gardner-Stephen, 2012).

Klasifikasi teks otomatis telah dianggap sebagai metode penting untuk mengelola dan memproses sejumlah besar dokumen dalam bentuk digital yang tersebar luas dan terus meningkat. Secara umum, klasifikasi teks memainkan peran penting dalam ekstraksi informasi, peringkasan, pencarian teks, dan tanya jawab (Ikonomakis, 2005). Namun memakan waktu dan tenaga bagi manusia untuk membaca ulang dan secara tepat mengkategorikan sebuah artikel secara manual. Upaya untuk mengatasi tantangan ini, studi klasifikasi dokumen otomatis mendapatkan lebih banyak minat dalam penelitian *text mining* baru-baru ini (Ting, Ip and Tsang, 2011).

Secara umum, pengolahan kategorisasi teks otomatis melibatkan dua masalah penting yang pertama adalah ekstraksi istilah fitur yang menjadi kata kunci yang efektif dalam tahap pelatihan, dan yang kedua adalah klasifikasi sebenarnya dari dokumen menggunakan istilah – istilah fitur dalam tahap uji coba (Suzuki and Hirasawa, 2007). Metode klasifikasi teks yang telah ada mungkin tidak cocok untuk menganalisis teks pendek karena tidak mempertimbangkan sifat khusus mereka. Sebagai contoh, sulit untuk menganalisis teks dengan banyak kata, tapi kita bisa dengan mudah mengetahui detail dalam pelatihan / memprediksi teks

pendek. Tidak seperti teks umum seperti halaman *web* dan email, teks pendek dalam korpus mungkin memiliki panjang yang sama dan kata - kata dalam teks singkat sering berbeda (Yu, Ho, Juan and Lin, 2013).

Teknik yang digunakan untuk klasifikasi teks yang bekerja untuk dataset dengan dokumen besar seperti karya ilmiah tidak bekerja dengan baik untuk dokumen dalam corpus dengan pelatihan yang singkat. Kehilangan kinerja dapat dikaitkan dengan tanda kelemahan dari konsep yang dimodelkan karena ukuran pendek dari teks (Healy, Delany and Zamolotskikh, 2004).

Penelitian terdahulu tentang klasifikasi teks untuk *Short Message Service* (SMS) masih sulit untuk diselesaikan oleh komputer tetapi menggunakan *machine learning* memberikan pendekatan yang menjanjikan untuk melakukan klasifikasi SMS menggunakan metode *term weighting* seperti metode *Term Frequency – Inverse Document Frequency* (TF-IDF) (Al-Talib and Hassan, 2013). Selain itu terdapat penelitian yang mengganti kata pendek dengan kata yang sebenarnya untuk meningkatkan akurasi dari klasifikasi SMS karena kata-kata pendek terbukti mengurangi tingkat akurasi dari klasifikasi teks SMS (Padhiyar, 2013). Klasifikasi teks menggunakan algoritma *Learning Vector Quantization* (LVQ) ini didasarkan pada Kohonen *Self-Organizing Map* (SOM) yang mampu mengatur ukuran besar dokumen koleksi menurut kesamaan tekstual. Algoritma LVQ memerlukan contoh pelatihan yang sedikit dan yang lebih cepat daripada metode klasifikasi lain (Azara, Mohammed, Fatayer, Tamer, El-Halees, 2012).

Dari latar belakang yang telah dikemukakan di atas, penulis terdorong untuk mengangkat topik “Klasifikasi Teks *Short Message Service* (SMS)”. Tujuan

mengangkat topik ini adalah untuk memberikan kontribusi dalam metode klasifikasi teks yang dikhususkan untuk teks singkat / pendek seperti SMS dan memiliki akurasi yang mampu melakukan klasifikasi SMS dengan tepat.

B. Perumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan, maka dapat diidentifikasi dan dirumuskan permasalahan umum yang akan dihadapi dalam melakukan klasifikasi SMS. Rumusan masalah tersebut antara lain sebagai berikut.

1. Bagaimana algoritma LVQ dalam melakukan klasifikasi SMS?
2. Bagaimana tingkat akurasi dalam klasifikasi SMS menggunakan algoritma LVQ untuk mengklasifikasikan SMS teks dengan kategori bencana, kejahatan, dan kecelakaan?

C. Batasan Masalah

Ruang lingkup permasalahan pada penelitian ini hanya akan dibatasi pada beberapa hal seperti di bawah ini :

1. Basis data untuk SMS didapatkan dari beberapa partisipan yang diminta secara khusus untuk membuat SMS dengan kategori tertentu.
2. Setiap SMS yang didapatkan dari partisipan akan digunakan sebagai basis data dalam melakukan pelatihan dan pengujian sistem dalam mengenali kategori SMS.
3. Kategori SMS yang akan digunakan hanya termasuk ke dalam 3 (tiga) jenis yaitu Kecelakaan, Kejahatan, dan Bencana.

D. Keaslian Penelitian

Penelitian yang berjudul “Klasifikasi Teks *Short Message Service*” ini belum pernah dilakukan oleh peneliti lain.

E. Manfaat yang diharapkan

Manfaat dari penelitian Klasifikasi Teks *Short Message Service* (SMS) dengan kategori kejahatan, bencana, dan kecelakaan ini adalah :

1. Mengetahui bagaimana melakukan klasifikasi teks pendek untuk SMS menggunakan algoritma Linear Vector Quantization (LVQ).
2. Mengetahui keakuratan algoritma LVQ untuk melakukan klasifikasi teks pendek SMS dengan kategori SMS kejahatan, bencana, dan kecelakaan. Salah satu bentuk penerapan mesin pembelajaran yang mampu membantu klasifikasi teks pendek seperti SMS dalam menangani jenis SMS yang berbeda yang berasal dari basis data SMS yang telah dikumpulkan.

F. Tujuan Penelitian

Tujuan dari penelitian Klasifikasi Teks *Short Message Service* (SMS) menggunakan algoritma LVQ ini adalah :

1. Mengetahui kemampuan algoritma LVQ untuk melakukan klasifikasi SMS dengan beberapa jenis kategori yang telah ditentukan sebelumnya.
2. Mengetahui presentasi keakuratan klasifikasi SMS menggunakan algoritma LVQ dengan kategori tindak kejahatan, kecelakaan, dan bencana.

G. Sistematika Penulisan

BAB I Pendahuluan

Bab pendahuluan memuat latar belakang, perumusan masalah, batasan masalah, keaslian penelitian, manfaat yang diharapkan, tujuan penelitian, dan sistematika penulisan.

BAB II Tinjauan Pustaka

Tinjauan pustaka memuat uraian sistematis tentang hasil-hasil penelitian yang didapat oleh peneliti terdahulu serta pustaka yang ada hubungannya dengan penelitian yang akan dilakukan. Dalam penyajian ini hendaknya ditunjukkan bahwa permasalahan yang akan diteliti belum terjawab atau belum terpecahkan secara memuaskan. Fakta yang dikemukakan diambil / dikutip dari sumber aslinya. Semua sumber yang dipakai harus disebutkan dengan mencantumkan nama penulis dan tahun penerbitannya. Tinjauan pustaka yang ada dalam tesis merupakan penyempurnaan dan perluasan proposal tesis.

BAB III Landasan Teori

Landasan teori dijabarkan dari tinjauan pustaka dan disusun sendiri oleh mahasiswa sebagai tuntunan untuk memecahkan masalah penelitian dan untuk merumuskan hipotesis. Landasan teori dapat berbentuk uraian kualitatif, model matematis, atau persamaan-persamaan yang langsung berkaitan dengan bidang ilmu yang diteliti.

BAB IV Metodologi Penelitian

Metodologi penelitian yang ada dalam tesis merupakan penyempurnaan dan perluasan proposal tesis. Pada cara penelitian terdapat uraian terinci tentang:

bahan atau materi penelitian, alat, langkah-langkah penelitian, analisis hasil dan kesulitan-kesulitan serta cara pemecahannya.

1. Bahan atau materi penelitian harus dinyatakan spesifikasinya dengan selengkap-lengkapnyanya. Hal ini perlu dikemukakan agar peneliti lain yang ingin menguji ulang penelitian itu tidak sampai salah langkah.
2. Alat yang dipergunakan untuk melaksanakan penelitian seperti misalnya kuesioner, daftar pertanyaan untuk wawancara, formulir pengamatan perlu diuraikan dengan jelas.
3. Langkah-langkah penelitian berupa uraian yang lengkap dan terinci tentang langkah-langkah yang telah diambil pada pelaksanaan penelitian, termasuk cara mengumpulkan data dan jenisnya.
4. Kesulitan-kesulitan yang timbul selama penelitian dan cara pemecahannya perlu sekali diungkapkan, agar para peneliti yang akan berkecimpung dalam bidang penelitian yang sejenis terhindar dari hal-hal yang tidak menyenangkan.

BAB V Hasil penelitian dan pembahasan

Bab ini memuat hasil penelitian dan pembahasan terpadu.

1. Hasil penelitian memuat uraian secara jelas dan tepat. Tabel, grafik dan gambar hendaknya ditemp
2. atkan sedekat-dekatnya dengan pembahasan.
3. Pembahasan berisi tentang analisis yang dilakukan terhadap hasil yang diperoleh, ditinjau secara utuh baik secara kualitatif, kuantitatif maupun normatif.

BAB VI Kesimpulan dan Saran

Kesimpulan dan saran harus dinyatakan secara terpisah.

1. Kesimpulan merupakan pernyataan singkat dan tepat yang dijabarkan dari hasil penelitian dan pembahasan. Apabila penelitian tersebut menggunakan hipotesis maka perlu dijelaskan kebenaran hipotesis.
2. Saran dibuat berdasarkan pengamatan dan pertimbangan penulis, ditujukan kepada para peneliti dalam bidang sejenis, yang ingin melanjutkan atau mengembangkan penelitian yang sudah diselesaikan. Saran dapat berisi keterbatasan dari penelitian yang telah dilakukan.